Can cohort data be treated as genuine panal data?

Nijman, T.E.; Verbeek, M.J.C.M.

Published in:
Empirical Economics: A quarterly journal of the Institute for Advanced Studies

Publication date:
1992

Empirical Economics (1992) 17:9–23

Can Cohort Data be Treated as Genuine Panel Data?

By M. Verbeek and T. Nijman[1,2]

Abstract: If repeated observations on the same individuals are not available it is not possible to capture unobserved individual characteristics in a linear model by using the standard fixed effects estimator. If large numbers of observations are available in each period one can use cohorts of individuals with common characteristics to achieve the same goal, as shown by Deaton (1985). It is tempting to analyze the observations on cohort averages as if they are observations on individuals which are observed in consecutive time periods. In this paper we analyze under which conditions this is a valid approach. Moreover, we consider the impact of the construction of the cohorts on the bias in the standard fixed effects estimator. Our results show that the effects of ignoring the fact that only a synthetic panel is available will be small if the cohort sizes are sufficiently large (100, 200 individuals) and if the true means within each cohort exhibit sufficient time variation.

1 Introduction

In recent years much attention is paid to the comparison of panel data with a single cross section or a series of independent cross sections (cf. Hsiao (1985)). In the context of a random effects model, for example, Nijman and Verbeek (1990) show that more efficient estimators of several functions of the parameters can be obtained from a series of cross sections than from a panel (with the same number of observations). On the other hand several authors have stressed the fact that panel data are not indispensible for the identification of many commonly estimated models (see, for example, Heckman and Robb (1985), Deaton (1985) and Moffitt (1990)). In this paper we pay attention to a regression model with in-

[1] The authors thank Bertrand Melenberg, Robert Moffitt, Guglielmo Weber, seminar participants at Texas A & M University, Rice University and the Conference on "The Econometrics of Panels and Pseudo Panels" (Venice, October 1990) and two anonymous referees for helpful comments. Rob Alessie and Pim Adang kindly provided the data. Financial support by the Royal Netherlands Academy of Arts and Sciences (K. N. A. W.) and the Netherlands Organization for Scientific Research (N. W. O.) is gratefully acknowledged.

[2] Marno Verbeek and Theo Nijman, Tilburg University, Department of Econometrics, P. O. Box 90153, 5000 LE Tilburg, The Netherlands.

0377–7332/92/1/9–23 $2.50

dividual effects that are correlated with the explanatory variables ("the fixed effects model"), and analyze the properties of the within estimator based on aggregated data on cohorts constructed from a series of independent cross sections. In this approach "similar" individuals are grouped in cohorts, after which the averages within these cohorts are treated as observations in a synthetic panel. Because the observed cohort aggregates are error-ridden measurements of the true cohort population values, Deaton (1985) proposes an errors-in-variables estimator which yields consistent estimators under fairly weak assumptions.

However, if the number of observations per cohort is large, it is tempting to ignore the errors-in-variables problem and to use standard software to handle the synthetic panel as if it were a genuine panel. This is what is usually done in empirical studies, see e.g., Browning, Deaton and Irish (1985) and Blundell, Browning and Meghir (1989). In this paper we analyze to what extent this is a valid approach. First, in Section 2, we present a general introduction and derive conditions for the consistency of the standard within estimator on the synthetic panel which ignores the measurement errors problem. In Sections 3 and 4 we derive expressions for the bias and the (estimated) variance of this estimator, respectively, if the conditions for consistency are not met. In Section 5 we consider the implications of our results for the estimation of Engel curves for food expenditures from Dutch monthly data. Finally, Section 6 concludes.

2 Estimation from Cohort Data

Consider the following linear model

$$y_{it} = x_{it}\beta + \theta_i + \varepsilon_{it}, \quad t = 1, \ldots, T \tag{1}$$

where i indexes individuals and t indexes time periods and suppose β is the parameter of interest. Throughout the paper we assume that $E\{\varepsilon_{it}|x_{js}\} = 0$ for all $s, t = 1, \ldots, T$ and all i, j. In each period, observations on N individuals are available. Throughout we assume that the data set is a series of independent cross sections.

In many applications the individual effects θ_i are likely to be correlated with the explanatory variables in x_{it} so that estimation procedures treating the θ_i as random drawings from some distribution lead to inconsistent estimators, unless the correlation is explicitly taken into account. When panel data are available this problem can be solved by treating the θ_i as fixed unknown parameters. Usually the fixed effects are eliminated before estimation, for example by a within or first difference transformation. Obviously, this strategy no longer applies if no repeated observations on the same individuals are available.

Deaton (1985) suggests the use of cohorts to obtain consistent estimators for β in (1) if repeated cross sections are available, even if the individual effects are correlated with the explanatory variables. Let us define C cohorts, which are groups of individuals sharing some common characteristics. These cohorts are defined in such a way that each individual is a member of exactly one cohort which is the same for all periods. For example, a particular cohort may consist of all male individuals born in 1945–1949. Aggregation of all observations to cohort level results in

$$\bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\theta}_{ct} + \bar{\varepsilon}_{ct} \ , \quad c = 1, \ldots, C \ ; \quad t = 1, \ldots, T \tag{2}$$

where $\bar{y}_{ct}$ and $\bar{x}_{ct}$ are the averages of all observed y_{it}'s and x_{it}'s in cohort c at time t. The resulting data set is a synthetic (or pseudo) panel with repeated observations on C cohorts over T periods. The main problem with the estimation of this model is that $\bar{\theta}_{ct}$ in (2) depends on t, is unobserved and is likely to be correlated with $\bar{x}_{ct}$. Therefore, treating the $\bar{\theta}_{ct}$ as random (and uncorrelated with the explanatory variables) is likely to lead to inconsistent estimators and treating them as fixed will result in an identification problem unless the variation of $\bar{\theta}_{ct}$ over t can be neglected. Intuitively, the latter will be the case if the number of observations within each cohort is large.

An alternative way to approach the problem is adopted by Deaton (1985), who considers the cohort population version of (1),

$$y^*_{ct} = x^*_{ct}\beta + \theta^*_c + \varepsilon^*_{ct} \ , \quad c = 1, \ldots, C \ ; \quad t = 1, \ldots, T \tag{3}$$

where the asterisks denote (unobservable) population cohort means and where θ^*_c is the cohort fixed effect, which is constant over time because population cohorts contain the same individuals in all periods. If the population cohort means would be observable, eq. (3) could be used to estimate β using standard procedures for a panel consisting of C cohorts observed in T periods. However, we can regard the observed cohort means $\bar{y}_{ct}$ and $\bar{x}_{ct}$ as error-ridden measurements of the true population cohort means y^*_{ct} and x^*_{ct}. Deaton (1985) assumes that the measurement errors in $\bar{y}_{ct}$ and $\bar{x}_{ct}$ are normally distributed with zero mean and independent of the true values y^*_{ct} and x^*_{ct}, in particular[3]

$$\begin{pmatrix} \bar{y}_{ct} \\ \bar{x}_{ct} \end{pmatrix} \sim N\left(\begin{pmatrix} y^*_{ct} \\ x^*_{ct} \end{pmatrix} \ ; \ \begin{pmatrix} \sigma_{00} & \sigma' \\ \sigma & \Sigma \end{pmatrix} \right) \ . \tag{4}$$

---

[3] Note that, contrary to Deaton, we do not include the cohort dummies in the vector of x's. These dummies are of course observed without error.

One way to estimate the parameter β in (3) is to analyze the model in (3) and (4) as a model with measurement errors. If the row vector of cohort dummies is denoted by d_c and the column vector of corresponding parameters is denoted by $\theta^* = (\theta_1^*, \ldots, \theta_C^*)'$, the errors-in-variables estimator (on the model in levels) proposed by Deaton (1985) is given by

$$\begin{pmatrix} \bar{\theta} \\ \bar{\beta} \end{pmatrix} = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' \bar{x}_{ct} \\ \bar{x}_{ct}' d_c & \bar{x}_{ct}' \bar{x}_{ct} - \hat{\Sigma} \end{pmatrix} \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' \bar{y}_{ct} \\ \bar{x}_{ct}' \bar{y}_{ct} - \hat{\sigma} \end{pmatrix} \right) \tag{5}$$

where $\hat{\Sigma}$ and $\hat{\sigma}$ are estimates of Σ and σ based on all individual observations. If the following assumption holds, the estimator $\bar{\beta}$ is consistent for β if the number of observations CT tends to infinity, while $\bar{\theta}$ is consistent for θ^* if the total number of observations per cohort (TN/C) tends to infinity.

Assumption 2.1 *The moments matrix of the population means of the explanatory variables*

$$\operatorname*{plim}_{CT \to \infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} \begin{pmatrix} d_c' d_c & d_c' \bar{x}_{ct} \\ \bar{x}_{ct}' d_c & \bar{x}_{ct}' \bar{x}_{ct} - \hat{\Sigma} \end{pmatrix} \tag{6}$$

is nonsingular.

If the number of observations per cohort is not too small, it is tempting to ignore the errors-in-variables problem and to estimate (2) assuming equality of population and sample means. The resulting estimator for β is the within estimator on the synthetic panel, $\hat{\beta}_W$, given by

$$\hat{\beta}_W = \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c) \right)^{-1} \left(\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c) \right), \tag{7}$$

where $\bar{x}_c$ is the time average of $\bar{x}_{ct}$, i.e. $\bar{x}_c = \frac{1}{T} \sum_{t=1}^{T} \bar{x}_{ct}$ and analogously for $\bar{y}_c$. Using (2) it is easy to show that $\hat{\beta}_W$ is unbiased if

$$E\{\bar{\theta}_{ct} - \bar{\theta}_c \mid \bar{x}_{ct} - \bar{x}_c\} = 0 \tag{8}$$

provided the following assumption holds.

Assumption 2.2 *The moments matrix of the observed cohort means of the explanatory variables*

$$\underset{CT\to\infty}{\text{plim}}\ \frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct}-\bar{x}_c)'(\bar{x}_{ct}-\bar{x}_c) \tag{9}$$

is nonsingular.

It is important to note that Assumption 2.2 is implied by Assumption 2.1 but that the converse is not true. Condition (8) will be satisfied if θ_i is independent of x_{it} (for all t) or if the averaged individual effects θ_i are constant over time ($\bar{\theta}_{ct}=\bar{\theta}_c$). If the number of observations per cohort, N/C, is large, one is tempted to assume that the latter condition holds. In the sequel of this paper we shall pay attention to the bias in the cohort within estimator $\hat{\beta}_W$ given the number of observations per cohort (N/C). Note that increasing the number of observations per cohort implies a decrease in the number of observations in the synthetic panel and thus an increase in the variance of the within estimator on the synthetic panel. Evidently, the optimal choice of the cohorts will depend on both its impact on the bias and its impact on the variance, which will be analyzed (for a simple model) in Sections 3 and 4, respectively.

A striking point from our results is that Deaton (1985)'s estimator has a nonexisting probability limit (for $CT\to\infty$), while $\hat{\beta}_W$ has a well-defined probability limit which may even equal the true value β if Assumption 2.2 is satisfied but Assumption 2.1 is not. We will return to this point in the next section.

3 The Effects of the Choice of Cohorts on the Bias

Our basic interest lies in the validity of the argument that "the number of observations per cohort is large enough to ignore the errors-in-variables problem" (cf., e.g., Browning, Deaton and Irish (1985)). We therefore concentrate on the case where the number of observations per cohort N/C is fixed. To simplify the analytical results we approximate the finite sample bias by the asymptotic bias for large C and large N. Numerical checks reveal that this approximation is accurate if C is not too small (10–20). In this section we will derive the bias in $\hat{\beta}_W$ for the special case of a linear model with only one explanatory variable,

$$y_{it}=\beta x_{it}+\theta_i+\varepsilon_{it} \tag{10}$$

where x_{it} is a scalar variable. Following Chamberlain (1984), we assume that the dependence of x_{it} and θ_i can be characterized as follows.

Assumption 3.1 *The individual effects θ_i are correlated with the x's in the following way*

$$\theta_i = \lambda \bar{x}_i + \xi_i \tag{11}$$

where $E\{\xi_i | x_{it}\} = 0$ *for all* $t = 1, \ldots, T$ *and* $V\{\xi_i\} = \sigma_\xi^2$.

Then, under Assumptions 2.2 and 3.1, $\lambda = 0$ is a sufficient condition for consistency of $\hat{\beta}_W$ as in that case the cohort effects $\bar{\theta}_{ct}$ in (2) are uncorrelated with the regressors. Cohorts are assumed to be constructed in the following way.

Assumption 3.2. *Cohorts are defined on the basis of an absolute continuous distributed variable* z *which is distributed independently across individuals with variance normalized to unity. Moreover, the cohorts are chosen such that the (unconditional) probability of being in a particular cohort is the same for all cohorts.*

According to this assumption the support of the density of z is split into C intervals with equal probability mass, implying that all cohorts have approximately the same number of members in the sample. In practice, the variable z may be based on more than one underlying variable. It should be noted that the choice of z (or the underlying variables) is restricted. First, z_i should be constant over time for each individual i because individuals cannot move from one cohort to another. Second, z_i should be observed for *all* individuals in the sample. The latter requirement rules out variables like "wage earnings in 1988" or "family size at January, 1st, 1990", because these variables are typically not observed for all individuals in the sample. In applications variables like date of birth or sex will be chosen to define the cohorts.

For Assumptions 2.1 and 2.2 to be satisfied it is required that the true cohort means vary over cohorts and/or over time. To model this, we assume that the correlation between x_{it} and z_i (on an individual level) is of the following form.

Assumption 3.3. *The regressor variable* x_{it} *is correlated with* z_i *in the following fashion*

$$x_{it} = \mu_t + \gamma_t z_i + v_{it} \tag{12}$$

where v_{it} *is uncorrelated with* z_i, *has expectation zero and (for the sake of simplicity) a constant variance* σ_v^2. *Its correlation over time is characterized by* $E\{v_{it} v_{is}\} = \varrho \sigma_v^2$ *if* $s \neq t$. *The* μ_t *are fixed (unknown) constants (fixed time effects).*

This assumption implies that v_{it} has the commonly assumed error components structure with an individual specific effect. The result can easily be generalized to, for example, the case where $E\{v_{it} v_{is}\} = \varrho_{|t-s|} \sigma_v^2 (s \neq t)$.

It can be shown (see Appendix) that under Assumptions 2.2, 3.1, 3.2 und 3.3 the asymptotic bias of the within estimator $\hat{\beta}_W$ is given by

$$\operatorname*{plim}_{C \to \infty} (\hat{\beta}_W - \beta) = \lambda \left[\frac{1 + (T-1)\varrho}{T} \right] \frac{\tau \omega_2}{\omega_1 + \tau \omega_2} = \delta \ , \tag{13}$$

where $\tau = (T-1)/T$, ω_2 is the measurement error variance in $\bar{x}_{ct}$, i.e.

$$\omega_2 = \operatorname*{plim}_{C\to\infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - x^*_{ct})^2 = n_c^{-1} \sigma_v^2 \,, \tag{14}$$

with n_c the number of individuals[4] in each cohort (N/C), and where ω_1 is the true within cohort variance[5]

$$\omega_1 = \lim_{C\to\infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (x^*_{ct} - \bar{x}^*_c)^2$$

$$= \frac{1}{T} \sum_{t=1}^{T} \left(\mu_t - \frac{1}{T} \sum_{s=1}^{T} \mu_s \right)^2 + \frac{1}{T} \sum_{t=1}^{T} \left(\gamma_t - \frac{1}{T} \sum_{s=1}^{T} \gamma_s \right)^2 \tag{15}$$

with $\bar{x}^*_c = \frac{1}{T} \sum_{t=1}^{T} x^*_{ct}$.

Under Assumption 3.3 it can be easily checked that Assumption 2.1 implies that $\omega_1 > 0$, while Assumption 2.2 implies that $\omega_1 + \tau\omega_2 > 0$. Note that $\omega_1 > 0$ requires that μ_t or γ_t vary with t. If this is not the case the probability limit of Deaton's errors-in-variables estimator does not exist, while the bias in the within estimator is maximal, i.e.

$$\operatorname*{plim}_{C\to\infty} (\hat{\beta}_W - \beta) = \lambda \left[\frac{1 + (T-1)\varrho}{T} \right] = \delta_{\max} \,, \tag{16}$$

which is independent of the cohort sizes. The choice of larger cohorts (decreasing ω_2) will reduce the bias if $\omega_1 > 0$ only. Because ω_2 is a decreasing function of n_c the bias in the within estimator is smallest if the number of observations in each cohort is as large as possible.

If ω_1/σ_v^2 is not too small the actual bias will be much smaller than the maximal bias if n_c is fairly large. Consider, as an example, the case where $\omega_1/\sigma_v^2 = 0.5$. Then one can easily compute that the bias will be less than 2% of the maximal bias if the cohorts have 100 or more observed members each. If ω_1/σ_v^2 is only 0.05 the bias will at most 17% of the maximal bias for cohort sizes of 100 or more. If these values are relevant for practical situations, this finding

---

[4] If cohort sizes are unequal the observations should be reweighted first by the square root of the cohort size, as in Deaton (1985).

[5] The true cohort means are treated here as fixed but unknown constants.

more or less justifies the fact that in most empirical studies (see, e.g., Browning, Deaton and Irish (1985) or Blundell, Browning and Meghir (1989); the measurement errors are ignored and the standard within estimator is used. It is important to note that cohort sizes may be chosen smaller if the cohort identifying variable is chosen in such a way that the true within cohorts variance is large relative to σ_v^2.

4 The Effects of the Choice of Cohorts on the Variance

In the previous section we have shown that the bias in the within estimator from the synthetic panel may be small if the number of observations per cohort is sufficiently large. However, an increase in the number of observations per cohort implies a decrease in the number of observations in the synthetic panel (CT) and – consequently – an increase in the variance of $\hat{\beta}_W$. In this section we will analyze the impact of the choice on the number of cohorts on this variance in more detail. Moreover, we show that the difference between the true variance of $\hat{\beta}_W$ and the probability limit of its routinely estimated variance is a function of the bias only.

The asymptotic variance of $\hat{\beta}_W$ can be written as

$$V\{\hat{\beta}_W\} = \frac{1}{CT}(\omega_1 + \tau\omega_2)^{-2} V^* \tag{17}$$

where

$$V^* = \lim_{C\to\infty} V\left\{\frac{1}{\sqrt{CT}} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{\theta}_{ct} - \bar{\theta}_c + \bar{\varepsilon}_{ct} - \bar{\varepsilon}_c)\right\} . \tag{18}$$

It should be noted that the expression within curved braces in (18) does not have expectation zero, because of the inconsistency of the estimator (if $\lambda \neq 0$). Moreover, the summations over c and t are neither summations over independently nor identically distributed variables. This complicates elaboration of the expression in (18). In the Appendix it is shown that under the additional assumption that $\bar{x}_{ct}, \bar{\theta}_{ct}$ and $\bar{\varepsilon}_{ct}$ are normally distributed, the variance of $\hat{\beta}_W$ is given by

$$V\{\hat{\beta}_W\} = \frac{1}{CT}[(\sigma_\theta^2 + \sigma_\varepsilon^2 n_c^{-1})(\omega_1 + \tau\omega_2)^{-1} + \delta^2] \tag{19}$$

where δ is the asymptotic bias of the within estimator defined in (13), and

$$\sigma_\theta^2 = \sigma_\xi^2 n_c^{-1} + \lambda^2 \left[\frac{1+(T-1)\varrho}{T} \right] \omega_2 \ , \tag{20}$$

which is the variance of $\bar{\theta}_{ct} - \theta_c^*$.

An increase in the cohort sizes n_c influences the variance of the within estimator $\hat{\beta}_W$ in two ways. First, the measurement error variance ω_2 and the equation error variance $\sigma_\theta^2 + \sigma_\varepsilon^2 n_c^{-1}$ are reduced. Second, the total number of observations CT is decreased. The latter effect is dominant, so an increase in n_c will cause a decrease in the variance of the within estimator on the synthetic panel. We will present some numerical results in the next section.

If standard software is used to compute $\hat{\beta}_W$, the routinely computed estimator of the variance,

$$\hat{V}\{\hat{\beta}_W\} = \hat{\sigma}^2 \left[\sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)^2 \right]^{-1} \tag{21}$$

will not be consistent for $V\{\hat{\beta}_W\}$ in (19). In general, it converges to

$$\bar{V}\{\hat{\beta}_W\} = \bar{\sigma}^2 \frac{1}{CT} (\omega_1 + \tau\omega_2)^{-1} \tag{22}$$

where

$$\bar{\sigma}^2 = \operatorname*{plim}_{C \to \infty} \hat{\sigma}^2 = \sigma_\varepsilon^2 n_c^{-1} + \sigma_\theta^2 - \delta^2 (\omega_1 + \tau\omega_2)^{-1} \ , \tag{23}$$

which is an underestimation of the true error variance $(\sigma_\xi^2 + \sigma_\varepsilon^2) n_c^{-1}$. Using (23) the probability limit of the estimated variance of $\hat{\beta}_W$ can be written as

$$\bar{V}\{\hat{\beta}_W\} = \frac{1}{CT} [(\sigma_\theta^2 + \sigma_\varepsilon^2 n_c^{-1})(\omega_1 + \tau\omega_2)^{-1} - \delta^2] \ . \tag{24}$$

As will be clear from the formulae above, the difference between the true variance and the probability limit of the estimated variance equals $2\delta^2/CT$ so it will be small if the bias δ is small.

5 An Empirical Illustration

In this section we consider the implications of the results in the previous sections on the estimation of Engel curves for food expenditures of Dutch households. We use a monthly panel data set to analyze what the properties of the within estimator on a synthetic panel would have been if one would analyze a series of repeated cross sections instead of a panel. The data used are the 367 complete monthly observations for 1986 in the so-called Expenditure Index Panel conducted by INTOMART, a marketing research agency in the Netherlands.

The model which is analyzed is the Engel curve for consumer expenditures on food,

$$w_{it} = \beta \log x_{it} + \theta_i + \varepsilon_{it} \ , \quad t = 1, \ldots, 12 \ , \tag{25}$$

where w_{it} is the budget share of food (in total expenditures on non-durables) and $\log x_{it}$ is the natural logarithm of total expenditures on non-durables. The individual effects θ_i reflect the influence of household specific characteristics (age, education, family size, etcetera) that are constant over the sample period (12 months). Obviously, these variables are likely to be correlated with total expenditures on non-durables and a fixed effects treatment of the θ_i is desired. As in the previous sections we shall impose Assumption 3.1,

$$\theta_i = \lambda \overline{\log x_i} + \xi_i \ . \tag{26}$$

The construction of the cohorts will be based on the data of birth of the head of the household, as in many applied studies. Because the relationship between age and total expenditures is likely to be nonlinear we choose the cohort identifying variable z_i as a quadratic function of the deviation of individual i's date of birth from the average date of birth in the sample (in years and months). The variance of z_i is normalized to one. Under Assumption 3.3 it holds that

$$\log x_{it} = \mu_t + \gamma_t z_i + v_{it} \ . \tag{27}$$

Using the 367 household observations of the balanced sub-panel, we easily obtain consistent estimates of the model parameters using ordinary least squares, which are given in Table 1. All estimated γ_t's are negative implying that (in a given period) total expenditures on non-durables are maximal at the average age of 49.2. Although all γ_t's and μ_t's differ significantly from zero, the variation in the γ_t's and μ_t's (reflected in $\omega_1 = 0.00681$) is relatively small in comparison with the estimated variance of v_{it}. Although the dependence of age and total expen-

Table 1. Parameter estimates based on 367 observations from the balanced sub-panel (standard errors – if computed – in parentheses)

β	−0.188 (0.006)	μ_1	12.235 (0.041)	γ_1	−0.147 (0.028)
λ	0.110 (0.007)	μ_2	12.085 (0.041)	γ_2	−0.132 (0.028)
σ_ζ	0.105	μ_3	12.202 (0.037)	γ_3	−0.164 (0.026)
σ_ε	0.072	μ_4	12.238 (0.041)	γ_4	−0.150 (0.028)
σ_v^2	0.305	μ_5	12.270 (0.043)	γ_5	−0.170 (0.030)
ϱ	0.634	μ_6	12.165 (0.041)	γ_6	−0.156 (0.028)
		μ_7	12.161 (0.046)	γ_7	−0.156 (0.022)
ω_1	0.00681	μ_8	12.152 (0.042)	γ_8	−0.139 (0.029)
		μ_9	12.180 (0.039)	γ_6	−0.154 (0.027)
		μ_{10}	12.328 (0.042)	γ_{10}	−0.162 (0.029)
		μ_{11}	12.224 (0.043)	γ_{11}	−0.181 (0.030)
		μ_{12}	12.385 (0.048)	γ_{12}	−0.233 (0.033)

ditures is significantly large, there does not seem to be much time variation in this dependence. Particularly for Deaton's errors-in-variables estimator this is something to worry about because its variance is inversely related with ω_1.

Before we discuss the consequences of these parameter values, we present some specification tests. First, we shall test the functional form of (25) by testing whether x_{it} (total expenditures on non-durables) should be included in (25). Subsequently we do the same for the triple x_{it}, x_{it}^2 and $\sqrt{x_{it}}$. This results in values for the Lagrange Multiplier test statistics of 2.75 and 7.83, respectively. Comparing these numbers with the critical values of a Chi-square distribution with one and three degrees of freedom, respectively, we do not take them as evidence against the null. Furthermore, we test Assumption 3.3, in particular the structure of the variance covariance matrix of v_{it}. We perform the (pseudo) LM test against first order autocorrelation, as discussed in Nijman and Verbeek (1990, Appendix), which yields a value of 0.057, clearly implying that we cannot reject our null hypothesis. Apparently, the error components structure imposed on v_{it} fits the data very well. In summary, we may conclude that our model is not evidently in conflict with the data.

From (16) we immediately obtain that the maximum bias in the within estimator based on cohort data over 12 periods equals 0.0731, which is 39% of the (estimated) true value. Given our choice of the cohort identifying variable it is possible to eliminate some of this bias by choosing large cohorts. This is illustrated in Table 2, where the theoretical biases in the within estimator are given for several cohort sizes. Note that the bias decreases slowly with the cohort size. In the table also the probability limit of the estimated standard error is given [based on (22)] and the true standard error [based on (19)]. Both are based on the asymptotic distribution. Although the bias is substantial the differences in these two standard errors are fairly small. Note that both standard errors increase if the cohort sizes are increased, which is caused by the fact that the number of (cohort)

Table 2. Bias in the standard within estimator $\hat{\beta}_W$, plim of estimated standard error and true standard error

n_c	bias (absolute)	bias (in %)	plim est. st. error/$\sqrt{N}$	true st. error/$\sqrt{N}$
2	0.0695	37.0	0.099	0.124
5	0.0650	34.6	0.152	0.171
10	0.0586	31.2	0.205	0.220
25	0.0453	24.1	0.287	0.298
50	0.0329	17.5	0.348	0.356
75	0.0258	13.7	0.379	0.386
100	0.0212	11.3	0.398	0.404
150	0.0157	8.3	0.420	0.424
200	0.0124	6.6	0.433	0.436

observations decreases if the cohort sizes are increased. Although there is the counteracting effect that the observations are more precise (contain less measurement error) if the cohort sizes are large, this effect is almost negligible.

6 Concluding Remarks

In this paper we analyzed the validity of treating cohort data as genuine panel data. Because the observed cohort averages are error-ridden measurements of the true cohort means, in general errors-in-variables estimators are required to obtain consistent estimators. If the individual effects and the explanatory variables in the model are correlated, a bias will occur in the standard fixed effects estimator, which will only be small if the number of observations in each cohort is large and if the time variation in the true cohort means is sufficiently large. To illustrate this we used genuine panel data on consumer expenditures to calibrate the possible magnitude of bias from using the synthetic panel data. The results show that in practice fairly large cohort sizes (100, 200 individuals) are needed to validly ignore the cohort nature of the data.

Appendix. Some Technical Details

In this appendix we sketch the derivation of (13) and (19). Using (12) we can write for the observed cohort means (in an obvious notation)

$$\bar{x}_{ct} = \mu_t + \gamma_t \bar{z}_{ct} + \bar{v}_{ct} = \mu_t + \gamma_t z_c^* + \bar{v}_{ct}^* = x_{ct}^* + \bar{v}_{ct}^* \tag{28}$$

where

$$z_c^* = E\{z_i | i \text{ is a member of cohort } c\} \tag{29}$$

and

$$\bar{v}_{ct}^* = \bar{v}_{ct} + \gamma_t(\bar{z}_{ct} - z_c^*) \ . \tag{30}$$

Furthermore, it follows from Assumption 3.1 for the aggregated individual effects $\bar{\theta}_{ct}$ that

$$\bar{\theta}_{ct} = \lambda \frac{1}{T}(\bar{x}'_{c1} + \bar{x}'_{c2} + \ldots + \bar{x}'_{cT}) + \bar{\xi}_{ct} \ , \tag{31}$$

where $\bar{x}'_{cs}$ is the average x-value in period s of all individuals observed in period t in cohort c. Notice that $\bar{x}'_{cs}$ is also an error-ridden measurement of x_{cs}^*, with the same properties as $\bar{x}_{ct}$ except that it is not observed. To be able to derive the probability limit of $\hat{\beta}_W$ we need expressions for the following probability limits

$$\operatorname*{plim}_{C\to\infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)^2 \tag{32}$$

and

$$\operatorname*{plim}_{C\to\infty} \frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)(\bar{\theta}_{ct} - \bar{\theta}_c) \ . \tag{33}$$

For the evaluation of (32) we use that[6]

$$E\left\{\frac{1}{CT} \sum_{c=1}^{C} \sum_{t=1}^{T} (\bar{x}_{ct} - \bar{x}_c)^2\right\}$$

$$= \frac{1}{T} \sum_{t=1}^{T} \left(\mu_t - \frac{1}{T} \sum_{s=1}^{T} \mu_s\right)^2 + \frac{1}{T} \sum_{t=1}^{T} \left(\gamma_t - \frac{1}{T} \sum_{s=1}^{T} \gamma_s\right)^2 \left(\frac{1}{C} \sum_{c=1}^{C} z_c^{*2}\right)$$

---

6 Convergence follows from applying Chebychev's weak law of large numbers.

$$+\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}E\left\{\left(\bar{v}_{ct}-\frac{1}{T}\sum_{s=1}^{T}\bar{v}_{cs}\right)^2\right\}$$

$$+\tau n_c^{-1}\frac{1}{T}\sum_{t=1}^{T}\gamma_t^2\frac{1}{C}\sum_{c=1}^{C}V\{z_i|i \text{ in cohort } c\}\ , \tag{34}$$

where $V\{z_i|i \text{ in cohort } c\}$ is the variance of z_i within cohort c. Because the total variance of z equals unity, increasing the number of cohorts implies that the distribution of z_c^* more and more resembles the distribution of z_i. Thus, the variance of z between the C cohorts satisfies

$$\lim_{C\to\infty}\frac{1}{C}\sum_{c=1}^{C}z_c^{*2}=1 \tag{35}$$

while

$$\lim_{C\to\infty}\frac{1}{C}\sum_{c=1}^{C}V\{z_i|i \text{ in cohort } c\}=1-\lim_{C\to\infty}\frac{1}{C}\sum_{c=1}^{C}z_c^{*2}=0\ . \tag{36}$$

Using that Assumptions 3.2 and 3.3 imply

$$\lim_{C\to\infty}E\left\{\frac{1}{CT}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{cs}^{t}-x_{cs}^{*})(\bar{x}_{cj}^{t}-x_{cj}^{*})\right\}=\varrho\omega_2,\quad j\neq s\ , \tag{37}$$

one can easily derive expressions for (32) and (33) to prove (13).

To derive the variance of $\hat{\beta}_W$ we have to elaborate (18). Under the normality assumption of $\bar{x}_{ct}, \bar{\theta}_{ct}$ and $\bar{\varepsilon}_{ct}$ the required fourth order moments can be written as functions of second order moments. In particular,

$$V\left\{\frac{1}{\sqrt{CT}}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct}-\bar{x}_c)(\bar{\theta}_{ct}-\bar{\theta}_c+\bar{\varepsilon}_{ct}+\bar{\varepsilon}_c)\right\}$$

$$=\frac{1}{CT}\sum_{c,d=1}^{C}\sum_{s,t=1}^{T}\left[E\{\tilde{x}_{ct}\tilde{x}_{ds}\}(E\{\bar{\theta}_{ct}\bar{\theta}_{ds}\}+E\{\bar{\varepsilon}_{ct}\bar{\varepsilon}_{ds}\})+E\{\tilde{x}_{ct}\bar{\theta}_{ds}\}E\{\tilde{x}_{ds}\bar{\theta}_{ct}\}\right]\ , \tag{38}$$

where $\tilde{x}_{ct}=\bar{x}_{ct}-\bar{x}_c$ and analogously for the other variables. Using straightforward algebra one can derive the following equalities.

$$E\{\bar{\theta}_{ct}\bar{\theta}_{ds}\} = \delta_{dc}\left(\delta_{ts} - \frac{1}{T}\right)(\sigma_\xi^2 n_c^{-1} + \lambda^2[T^{-1} + \tau\varrho]\omega_2) \tag{39}$$

and,

$$E\{\tilde{x}_{ct}\bar{\theta}_{ds}\} = \delta_{dc}\left(\delta_{ts} - \frac{1}{T}\right)\lambda[T^{-1} + \tau\varrho]\omega_2\ , \tag{40}$$

where δ_{ij} is the Kronecker δ satisfying $\delta_{ij} = 1$ if $i = j$, 0 otherwise. Using these equalities the variance V^* is readily obtained.

References

Blundell R, Browning M, Meghir C (1989) A microeconometric model of intertemporal substitution and consumer demand. Discussion Paper in Economics 89–11, University College London

Browning M, Deaton A, Irish M (1985) A profitable approach to labor supply and commodity demands over the life cycle. Econometrica 53:503–543

Chamberlain G (1984) Panel data. In: Griliches Z, Intriligator MD (eds) Handbook of Econometrics Vol II, North Holland, Amsterdam

Deaton A (1985) Panel data from time series of cross-sections. Journal of Econometrics 30:109–126

Heckman JJ, Robb R (1985) Alternative models for evaluating the impact of interventions: an overview. Journal of Econometrics 30:239–267

Hsiao C (1985) Benefits and limitations of panel data. Econometric Reviews 4:121–174

Moffitt R (1990) Estimating dynamic models with a time series of repeated cross sections, mimeo. Brown University, Providence RI

Nijman ThE, Verbeek M (1990) Estimation of time dependent parameters in linear models using cross sections, panels or both. Journal of Econometrics 46:333–346